SYSTEM FOR ASSEMBLING LARGE DATABASES THROUGH INFORMATION EXTRACTED FROM TEXT SOURCES

37 C.F.R. 1.71 AUTHORIZATION

37 C.F.R. 1.96 MICROFICHE (5 SHEETS/290 FRAMES) APPENDIX

Attached in microfiche is the MASTER.XLS spreadsheet specifying the required content of the Structured entity relationship model and all Entities, Links, Attributes and their values and the SOL Inserts used to set up the Research Strategy, Strategy Groups and Strategy Rules as used in the preferred embodiment of the invention. Furthermore, Appendix contains the Grammar Rules, the Application Data Tables, Application Definitions, and Research Rules as used in the prefered embodiment as specified below.

FIELD AND BACKGROUND OF THE INVENTION

The invention relates generally to natural language processing systems and information extraction processes that involve the extraction of information from source documents. An information extraction process should be distinguished from two other natural language processes: text or informational retrieval processes and text understanding processes. Text or informational retrieval processes typically identify documents from a library of documents by matching of words or phrases contained within the documents. Text understanding processes aim to interpret the complete meaning of entire texts, including the text's subtle nuances of meaning and complexities of language.

Traditional information extraction processes are usually implemented on a programmed general purpose computer. The process looks for certain information in the text, extracts the information, and organizes the information into a database records. The database created is usually stored in a searchable format, such a structured relational database or an object-oriented structured database, which can be accessed, research, and analyzed by computer-implemented database research systems.

In "The Generic Information Extraction System", Proceedings of the Fifth Message Understanding Conference (MUC-5), 1993, by J. R. Hobbs describes a generic information extraction system in ten steps. First, a text zoner, turns a text into a set of text segments, then a pre-processor turns a text or text segment into a sequence of sentences, each of which is a sequence of lexical terms, where a lexical item is a word together with its lexical attributes. Third, a filter turns a set of sentences into a smaller set of sentences by filtering out the irrelevant sentences. Fourth, a preparser takes a sequence of lexical items and tries to identify various determinable small-scale structures. Fifth, a parser produces a set of parse tree fragments from the sequence of lexical terms and small-scale structures. Sixth, a fragment combiner combines the fragments into a parse tree or logical form. Seventh, a semantic interpreter generates a semantic structure from the parse tree or logical form. Eighth, a lexical disambiguator replaces general or ambiguous predicates in the semantic structure with specific and unambiguous predicates. Ninth, a discourse or conference resolution processor turns the tree-like semantic structure into a network like structure by identifying different descriptions of the same entity in different parts of the text. Finally, a template generator derives the output template from the final semantic structure. Accordingly in "Tasks, Domains, and Languages", Proceedings of the Fifth Message Understanding Conference (MUC-5), 1993, by B. Onyshkevych et al. the tasks requested of the information extraction systems evaluated at MUC-5 focused on the systems' ability to automatically fill one singular object-oriented template with information extracted from a source of free text.

However, the generic information extraction process described above only inputs the extracted information into the database, or template, in the last step of the process, and does not address the problem of compiling or aggregating a large and comprehensive database from a plurality of source documents.

In addition, the information extraction processes do not address the problem of compiling or aggregating information extracted from both structured and unstructured source material, i.e. free text, forms.

Furthermore, the information extraction processes are not focused on how the information extracted will be used in the construction of a large database. It would be a desirable feature to have an information extraction system with the ability to assemble extracted information and to recognize any conflicts between the extracted information and the contents of an existing database. It would also be a desirable feature for this information extraction and assemblage system, or information indexing system, to fully analyze the extracted information in a comprehensive and intelligent manner to provide a full range of options and alternatives to the user to resolve any inconsistencies between the extracted information and the database under construction.

Another desirable feature would be for the system for the information indexing system to have the capacity of maintaining conflicts and fragments of incomplete information until they are resolved at a later date.

Accordingly, an information indexing system with the above features would have the ability to construct a database with a high degree of integrity from information extracted from a plurality of source documents.

SUMMARY OF THE INVENTION

The invention is an information indexing system which has the ability to create a database with a high degree of integrity from a plurality of text containing source documents. The invention is a system of combining information extracted from a plurality of different text containing source documents to produce a final databases with a high degree of integrity, that is a database built to contain information with both the maximum amount of corroboration possible, and maximum amount of cross-referencing for uncorroborated or ambiguous information. In the preferred embodiment of the invention, as described below, it is an additional feature of the invention that the information indexing system increases the efficiency of data collection and analysis in the preferred embodiment's application domain in criminal investigation procedures and analysis.

It is an additional feature of the invention that information extracted from a source document can be extracted from both the free text within a document and additional information presented in any structured format within the document. An example in the preferred embodiment's applica-